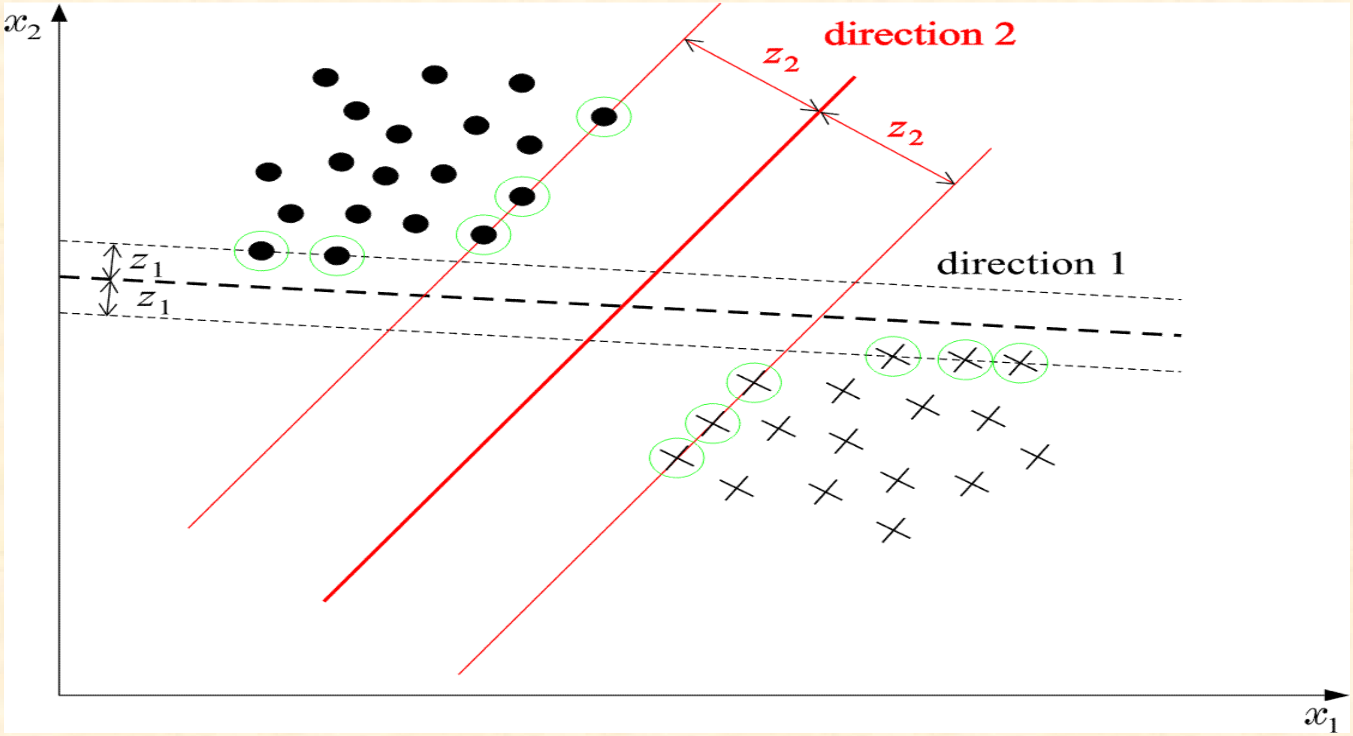
❖ ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΣΤΙΡΙΞΗΣ (SUPPORT VECTOR MACHINES)

- Ο στόχος: Δοθέντων δύο γραμμικώς διαχωρίσιμων κλάσεων, να σχεδιαστεί ο ταξινομητής

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0 = 0$$

που αφήνει το μέγιστο περιθώριο από τις δύο κλάσεις



- Περιθώριο: Κάθε υπερεπίπεδο χαρακτηρίζεται από
  - Την κατεύθυνση στο χώρο, δηλ., $\underline{w}$
  - Τη θέση του στο χώρο, δηλ., w_0
  - Για ΚΑΘΕ κατεύθυνση, $\underline{w}$
    επέλεξε το υπερεπίπεδο που αφήνει την ΙΔΙΑ απόσταση από τα πλησιέστερα σημεία από κάθε κλάση. Το περιθώριο είναι διπλάσιο αυτής της απόστασης.

- Η απόσταση ενός σημείου $\underline{\hat{x}}$ από ένα υπερεπίπεδο είναι

$$z_{\hat{x}} = \frac{g(\underline{\hat{x}})}{\|\underline{w}\|}$$

- Κλιμάκωσε τα, $\underline{w}, \underline{w}_0$, έτσι ώστε στα εγγύτερα από κάθε κλάση η συνάρτηση διάκρισης είναι ±1:

$$|g(\underline{x})| = 1 \quad \{g(\underline{x}) = +1 \text{ για } \omega_1 \text{ και } g(\underline{x}) = -1 \text{ για } \omega_2\}$$

- Έτσι το περιθώριο δίνεται από τη σχέση

$$\frac{1}{\|\underline{w}\|} + \frac{1}{\|\underline{w}\|} = \frac{2}{\|\underline{w}\|}$$

- Επίσης, ισχύουν τα ακόλουθα

$$\underline{w}^T \underline{x} + w_0 \geq 1 \;\; \forall \underline{x} \in \omega_1$$

$$\underline{w}^T \underline{x} + w_0 \leq -1 \;\; \forall \underline{x} \in \omega_2$$

- <u>SVM (γραμμικός) ταξινομητής</u>

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0$$

- Ελαχιστοποίησε

$$J(\underline{w}) = \frac{1}{2}\|\underline{w}\|^2$$

- Υπό τους περιορισμούς

$$y_i(\underline{w}^T \underline{x}_i + w_0) \geq 1,\ i = 1,2,\ldots,N$$

$$y_i = 1,\ \text{for } \underline{x}_i \in \omega_1,$$

$$y_i = -1,\ \text{for } \underline{x}_i \in \omega_2$$

- Αυτό γίνεται διότι ελαχιστοποιώντας το $\|\underline{w}\|$

το περιθώριο $\dfrac{2}{\|w\|}$ μεγιστοποιείται

- Η παραπάνω είναι μία διαδικασία τετραγωνικής βελτιστοποίησης (quadratic optimization task) υπό τον περιορισμό ενός συνόλου γραμμικών ανισοτήτων. Οι Karush-Kuhh-Tucker συνθήκες, δηλώνουν ότι το ελάχιστο ικανοποιεί τις συνθήκες:

  - (1) $\frac{\partial}{\partial \underline{w}} \mathrm{L}(\underline{w}, w_0, \underline{\lambda}) = 0$
  - (2) $\frac{\partial}{\partial w_0} L(\underline{w}, w_0, \underline{\lambda}) = 0$
  - (3) $\lambda_i \geq 0, i = 1,2,\ldots,N$
  - (4) $\lambda_i \left[y_i (\underline{w}^T \underline{x}_i + w_0) - 1 \right] = 0, i = 1,2,\ldots,N$

  - Όπου $L(.,.,.)$ είναι η συνάρτηση Lagrange

$$L(\underline{w},\ w_0,\ \underline{\lambda}) \equiv \frac{1}{2} \underline{w}^T \underline{w} - \sum_{i=1}^{N} \lambda_i [y_i (\underline{w}^T \underline{x}_i + w_0) - 1]$$

- Η λύση: από τα παραπάνω, προκύπτει ότι

  - $$\underline{w} = \sum_{i=1}^{N} \lambda_i y_i \underline{x}_i$$

  - $$\sum_{i=1}^{N} \lambda_i y_i = 0$$

- Σχόλια:
  - Οι πολ/στές Lagrange μπορεί να είναι είτε μηδέν είτε θετικοί. Έτσι,
    - $$\underline{w} = \sum_{i=1}^{N_s} \lambda_i y_i \underline{x}_i$$

      όπου $N_s \leq N$ αντιστοιχεί στους θετικούς πολ/στές Lagrange

    - Από τον περιορισμό (4) πιο πάνω, δηλ.,

      $$\lambda_i [y_i (\underline{w}^T \underline{x}_i + w_0) - 1] = 0, \quad i = 1,2,\ldots,N$$

      τα διανύσματα που συνεισφέρουν στο $\underline{w}$ ικανοποιούν

      $$\underline{w}^T \underline{x}_i + w_0 = \pm 1$$

– Τα διανύσματα αυτά καλούνται ΔΙΑΝΥΣΜΑΤΑ ΣΤΗΡΙΞΗΣ και είναι τα κοντινότερα διανύσματα, από κάθε κλάση, προς τον ταξινομητή.

– Μετά τον υπολογισμό του $\underline{w}$ το w_0 προσδιορίζεται από τις συνθήκες (4).

– Ο βέλτιστος ταξινομητής-υπερεπίπεδο μίας μηχανής διανυσματικής υποστήριξης είναι ΜΟΝΑΔΙΚΟΣ.

- Σχηματισμός Δυϊκού προβλήματος
  - Το πρόβλημα των SVM είναι διατυπωμένο ως ένα κυρτό πρόβλημα προγραμματισμού (convex programming problem), με
    - Κυρτή συνάρτηση κόστους
    - Κυρτή περιοχή εφικτών (feasible) λύσεων
  - Έτσι, η λύση μπορεί να προέλθει από το δυϊκό του πρόβλημα, δηλ.,
    - Μεγιστοποίησε $\underset{\underline{\lambda}}{} L(\underline{w}, w_0, \underline{\lambda})$
    - Υπό τους περιορισμούς

$$\underline{w} = \sum_{i=1}^{N} \lambda_i y_i \underline{x}_i$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

$$\underline{\lambda} \geq \underline{0}$$

- Συνδυάζοντας τα παραπάνω παίρνουμε
  - Μεγιστοποίησε την $(\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{ij}\lambda_i\lambda_j y_i y_j \underline{x}_i^T \underline{x}_j)$
  - Υπό τους περιορισμούς

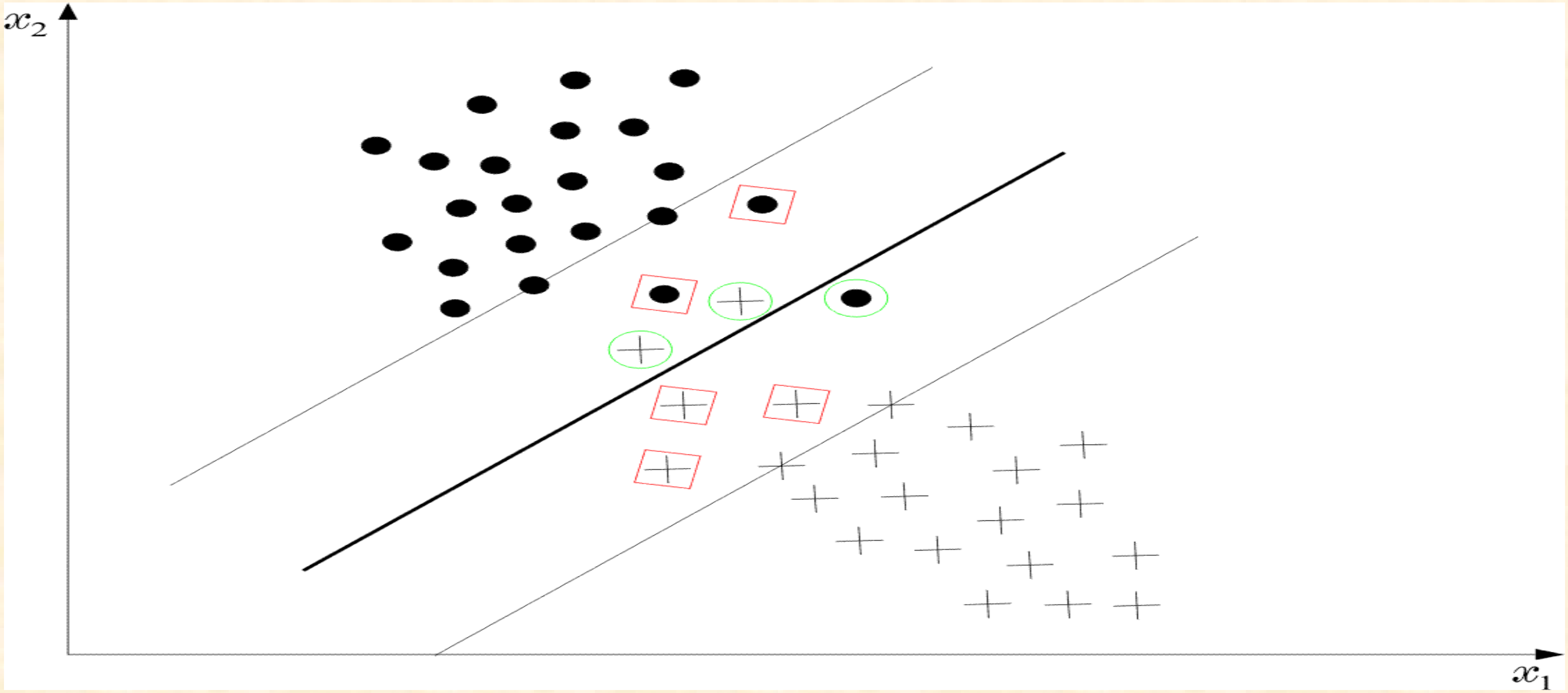
$$\sum_{i=1}^{N}\lambda_i y_i = 0$$

$$\underline{\lambda} \geq \underline{0}$$

➢ <u>Σχόλια</u>:

- Τα διανύσματα στήριξης εισέρχονται μέσω εσωτερικών γινομένων
- Παρότι η λύση, $\underline{w}$, είναι μοναδική, οι πολ/στές Lagrange <u>ΔΕΝ ΕΙΝΑΙ</u>.

➢ Non-Separable classes



Στην περίπτωση αυτή δεν υπάρχει υπερεπίπεδο, ώστε

$$\underline{w}^T \underline{x} + w_0 (><) 1, \ \forall \underline{x}$$

- Θυμηθείτε ότι το περιθώριο ορίζεται ως η απόσταση μεταξύ των δύο ακόλουθων υπερεπιπέδων

$$\underline{\boldsymbol{w}}^T \underline{\boldsymbol{x}} + \boldsymbol{w}_0 = 1$$

$$\text{και}$$

$$\underline{\boldsymbol{w}}^T \underline{\boldsymbol{x}} + \boldsymbol{w}_0 = -1$$

- Τα διανύσματα εκπαίδευσης ανήκουν σε μία από τρεις δυνατές κατηγορίες

  - A. Διανύσματα εκτός περιθωρίου που ταξινομούνται σωστά, δηλ.,

$$y_i(\underline{w}^T \underline{x} + w_0) > 1$$

  - B. Διανύσματα εντός περιθωρίου, που είναι σωστά ταξινομημένα, δηλ.,

$$0 \le y_i(\underline{w}^T \underline{x} + w_0) < 1$$

  - C. Διανύσματα λανθασμένα ταξινομημένα, i.e.

$$y_i(\underline{w}^T \underline{x} + w_0) < 0$$

- Όλες οι παραπάνω περιπτώσεις μπορούν να αναπαρασταθούν από

$$y_i(\underline{w}^T \underline{x} + w_0) \geq 1 - \xi_i$$

A. $\rightarrow \xi_i = 0$

B. $\rightarrow 0 < \xi_i \leq 1$

C. $\rightarrow 1 < \xi_i$

Οι ξ_i είναι γνωστές ως μεταβλητές χαλαρότητας (slack variables)

- Ο στόχος της βελτιστοποίησης είναι τώρα διττός
  - Μεγιστοποίηση περιθωρίου
  - Ελαχιστοποίηση του αριθμού των διανυσμάτων με $\xi_i > 0$

  Δηλαδή,

$$J(\underline{w}, w_0, \underline{\xi}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{N} I\left(\xi_i\right)$$

όπου C σταθερά και

$$I(\xi_i) = \begin{cases} 1 & \xi_i > 0 \\ 0 & \xi_i = 0 \end{cases}$$

  - I(.) δεν είναι παραγωγίσιμη. Στην πράξη, χρησιμοποιούμε μία προσέγγιση
  - $J(\underline{w}, w_0, \underline{\xi}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{N} \xi_i$
  - Εργαζόμενοι όπως και προηγουμένως παίρνουμε

- <u>KKT συνθήκες</u>

$$(1)\quad \underline{w} = \sum_{i=1}^{N} \lambda_i y_i \underline{x}_i$$

$$(2)\quad \sum_{i=1}^{N} \lambda_i y_i = 0$$

$$(3)\quad C - \mu_i - \lambda_i = 0, i = 1,2,\ldots,N$$

$$(4)\quad \lambda_i [y_i(\underline{w}^T \underline{x}_i + w_0) - 1 + \xi_i] = 0, \quad i = 1,2,\ldots,N$$

$$(5)\quad \mu_i \xi_i = 0, \quad i = 1,2,\ldots,N$$

$$(6)\quad \mu_i, \lambda_i \geq 0, \quad i = 1,2,\ldots,N$$

- Το σχετικό δυϊκό πρόβλημα

Μεγιστοποίησε $_{\underline{\lambda}} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \underline{x}_i^T \underline{x}_j\right)$

υπό τις προϋποθέσεις

$$0 \le \lambda_i \le C,\ i = 1,2,\ldots,N$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

- **<u>Σχόλιο</u>**:

Η μόνη διαφορά με την περίπτωση των διαχωρίσιμων κλάσεων είναι η ύπαρξη του *C* στους περιορισμούς.